TITLE: Comberons from single cell transcriptomics in endothelial cells

Sergii Domanskyi[1], Alex Hakansson[2], Michelle Meng[2], Joshua S. Graff Zivin[3], Carlo Piermarocchi[1], Giovanni Paternostro[2], Napoleone Ferrara[4]

[1]Department of Physics and Astronomy, Michigan State University, East Lansing, MI, 48824, USA
[2]Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, 92037, USA
[3]School of Global Policy and Strategy and Department of Economics, University of California, San Diego 9500 Gilman Drive, MC 0519, La Jolla, CA 92093, USA
[4]Moores Cancer Center, University of California San Diego, La Jolla, CA 92093, USA

Corresponding authors:
domansk6@msu.edu
jgraffzivin@ucsd.edu
piermaro@msu.edu
giovanni@sbpdiscovery.org
nferrara@health.ucsd.edu

ABSTRACT

VEGF inhibitor drugs have been successful, especially in ophthalmology, but not all patients respond to them. Combinations of drugs are likely to be needed for a really effective therapy of angiogenesis-related diseases. In this paper we introduce a new concept, the comberon, a term named by analogy with the operon that refers to evolutionarily conserved combinations of co-expressed genes. These genes identify potential drug targets.

Our results show that single-cell gene expression data can help to identify a set of co-expressed endothelial cell receptors, conserved among species (mice and human) and enriched, within a network, of connections to up-regulated genes. This set does include VEGF receptors and includes several other receptors previously shown to play a role in angiogenesis. The conclusions are supported by multiple highly significant statistical tests from large datasets, including an independent validation set. This discovery has broad pharmacological and socio-economic implications, going beyond angiogenesis therapy.

INTRODUCTION

Angiogenesis is the development of new blood vessels. It is a physiological process required for embryonic development, adult vascular homeostasis and tissue repair (1). It is a fundamental process of human biology involved in a large number of diseases, including cancer, eye diseases, arthritis, psoriasis and many other ischemic, inflammatory, infectious, and immune disorders.

During tumor progression, the new vessels provide neoplastic tissues with nutrients and oxygen; in intraocular vascular disorders, growth of abnormal, leaky blood vessels may destroy the retina and lead to blindness (1, 2). Extensive efforts to dissect the molecular basis of angiogenesis and to identify therapeutic targets resulted in the discovery of key signaling pathways involved in vascular development and differentiation (1, 3). In particular, numerous studies have established the pivotal role of the VEGF (Vascular Endothelial Growth Factor) pathway in physiological and pathological angiogenesis and therapies targeting this pathway have achieved success in treatment of cancer and ocular disorders (4, 5). VEGF inhibitors are currently the main class of FDA approved anti-angiogenic drugs and they are used as a therapy for cancer and for eye diseases leading to blindness. One of us has provided fundamental contributions to the elucidation of the role of VEGF in angiogenesis and to the development of these drugs (5).

VEGF inhibitor drugs have been successful, especially in ophthalmology, but not all patients respond to them (1, 4, 5). This is not surprising because combinations of cytokines acting on multiple endothelial cell receptors regulate the normal biological control of angiogenesis. Combinations of drugs acting on these molecules are likely to be needed for a really effective therapy of angiogenesis-related diseases (1, 4, 5). The challenge is therefore to discover new drug combinations able to effectively control angiogenesis in a broad spectrum of diseases.

Several methods are currently used to discover novel drug combinations (6). Beyond empirical combinations of individually effective drugs, they include the use of drug screens, using brute force (that is, for small sets, exhaustive testing of all possible combinations) or search algorithms depending on the phenotypic response (7). They also include modeling of biological systems, based on statistical associations or on direct modeling of molecular interactions (8). These approaches can also be combined, but our knowledge of the details of *in vivo* biological systems is still incomplete and progress to the clinic is therefore slow and inefficient. We also know that gene products are co-expressed in sets that have specific functions and these might represent potential drug targets for drug combinations (9). Understanding the regulation and role of these co-expressed sets has been a key focus of biological investigations, starting from the discovery of the *lac* operon in bacteria by Jacobs and Monod (10). In this paper we introduce a new concept, the comberon, a term that refers to evolutionarily conserved combinations of co-expressed genes and that is named by analogy with the operon of Jacob and Monod. The precise quantitative definition of the comberons and the statistical support for their existence is presented in the Methods and Results sections.

The increasing number of single-cell RNA-seq datasets that provide information on organ and tissue-specific signaling and dedicated initiatives such as the Human Cell Atlas, the NIH Human Biomolecular Atlas Program and the EU Lifetime Initiative (11, 12) aiming to map all the cell types of the human body, offer a new source of data that can inform the understanding of co-expression at the single cell level. Co-expression at the single cell level gives a more fundamental understanding of biological regulation compared to bulk measurements.

Evolution by natural selection is the main organizing principle in biology (13). The evolutionary age and rate of change of genes can provide information about their functions and interactions (14). We expect that any functionally important co-expression of genes would be conserved during evolution. Considerations analogous to these have guided combinatorial interventions before, for example the finding by Shin Yamanaka and his collaborators that delivering the same four transcription factors can reprogram cells to a pluripotent state in both mice and humans (15).

Furthermore, while incomplete, network models of intracellular interactions are being developed using an increasing number of datasets and can contribute to the validation of new findings (8). Endothelial cells, the main targets of VEGF, are the key cell type in angiogenesis and their receptors integrate multiple signals to decide if angiogenesis should be activated (16, 17). We can identify receptors linked, within network models, to significantly up-regulated genes, and we call this measure network enrichment. This is a measure of statistical association, but, in some cases, it might represent downstream changes due to the signals transmitted by the receptors.

We therefore present here an integrated analysis of multiple single cell RNA-seq datasets, using co-expression, evolutionary conservation and network enrichment to answer the following questions:

- Can single cell data help us to identify a set of co-expressed endothelial cell receptors, conserved among species (mice and human) and enriched, within a pre-determined network, for connections to up-regulated genes?
- Does this set include VEGF receptors and other receptors previously shown to play a role in angiogenesis?
- Can this set be identified and validated from a joint analysis of endothelial cells from multiple organs and tissues?
- Are there tissue-specific sets of co-expressed receptors?

These questions are addressed by introducing a new software tool, DECNEO (DEndrogram of Co-expression with Evolutionary Conservation and Network Enrichment Ordering), that identifies general and tissue-specific receptor comberons in endothelial cells and provides visualization and statistical analysis.

METHODS

Data sources

Three sources of data were used in our analysis:

(1) The primary data source in our study was PanglaoDB, a web-database of scRNA-seq datasets of *Homo sapiens* and *Mus musculus* studies, that collects 22,382,464 cells from 236 studies (SRAs, or Sequence Read Archive submission

accessions) in 1368 samples (SRSs, or Sequence Read Archive sample accessions), spanning numerous tissues and cell types (see Table S5). After quality control Panglao DB provides data from 4,459,768 *Mus musculus* cells and 1,126,580 *Homo sapiens* cells. Each of the datasets were downloaded from NCBI SRA and mapped to the GRCm38/GRCh38 genome assembly by the authors of PanglaoDB (18). All cells in PanglaoDB are annotated with cell type labels by Alona (19), a web-based bioinformatics tool for scRNA-seq analysis. This cell type annotation is incorporated in the PanglaoDB database. In addition to the Alona annotation, we processed each of the datasets in PanglaoDB using a platform for cell type annotation that we have recently developed (20, 21), DCS 1.3.6.10.

(2) The second source of data was a dataset from Voigt et al. (22) containing *Homo sapiens* choroid scRNA-seq on 18621 cells in 8 samples enriched for endothelial cells. The dataset was downloaded from NCBI and re-mapped with CellRanger 4.0.0 and the GRCh38 genome assembly; each of the 8 samples has 4 paired-end reads sequencing runs (SRRs) that were merged into one. In our alternative implementation of DECNEO, used as a validation, we used the original Voigt et al. (22) data mapping, quality control, and cell type annotation of the choroid cells.

(3) We also obtained an independent dataset for validation, composed of 14 single cell mice gene expression studies, containing 71 samples and 65,057 endothelial cells (see Table S5). This independent dataset was obtained after the analysis of the main dataset had already been completed and did not affect any of the methodological choices.

Input data format

See Figure 1A for a graphical summary of the analysis pipeline. The method is designed to take two single-cell transcriptomics datasets corresponding to two different species, e.g. *Mus musculus* and *Homo sapiens*. Each of the input datasets has to be in a table format where gene identifiers are in rows and cells and samples identifiers are in columns. Where possible, gene identifiers should be converted to one species, e.g. *Homo sapiens*, using their homologs. Starting from the gene counts matrix, the gene expression values should be scaled so that the sum of values of each cell in the dataset is identical (we recommend a scaling factor of 10^5) and then log-transformed. Cell type and sample annotation is required for each cell in each dataset, assigning cells to at least two different classes, e.g. endothelial cells and non-endothelial cells. The preprocessing step described above is necessary for the analysis pipeline. The preprocessing step is not directly included in DECNEO, but DCS (20) Domanskyi, 2021 #2859}, can be used to normalize and annotate cells.

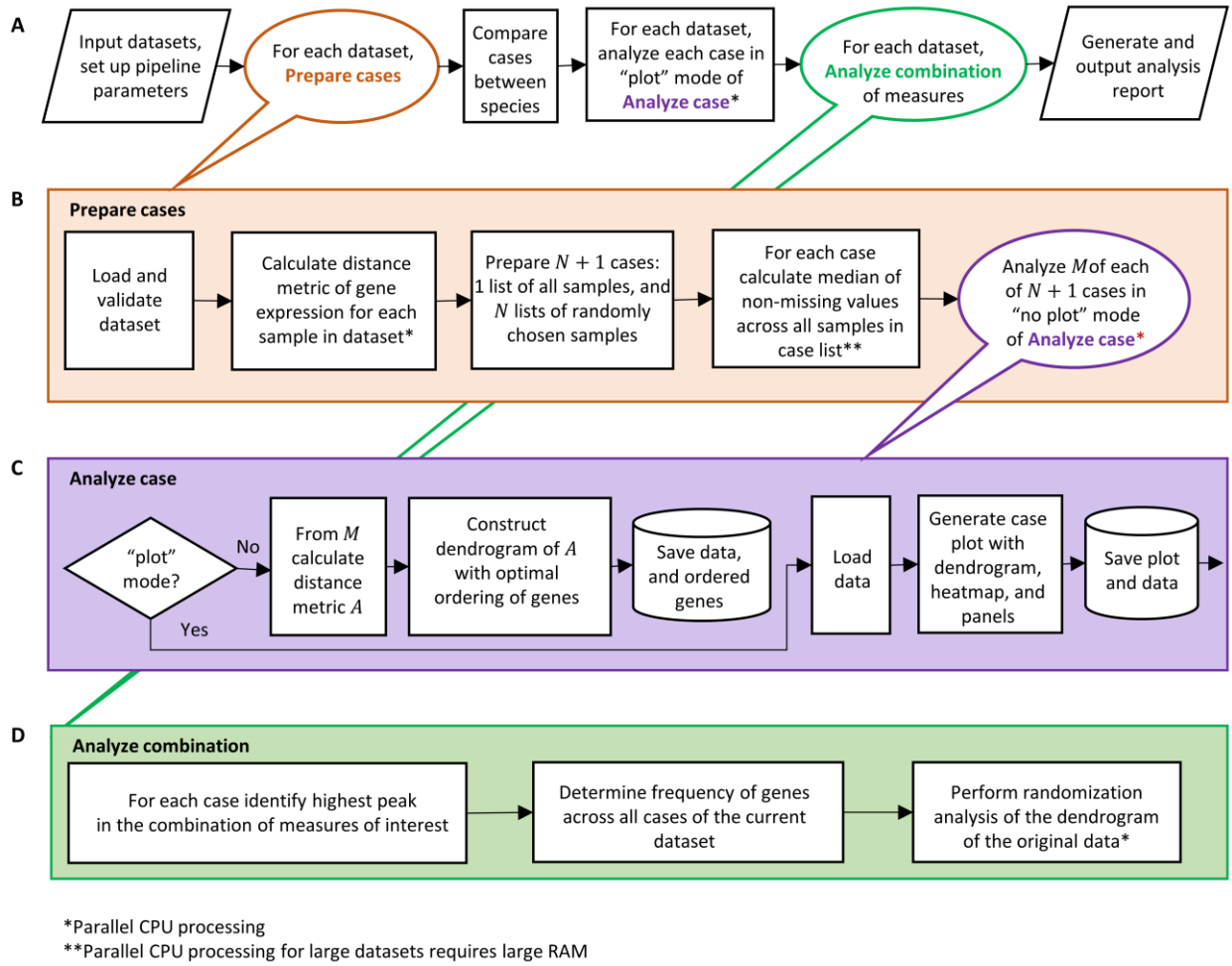

Figure 1. Summary of the analysis pipeline. Normalized single cell transcriptomics datasets are processed in steps depicted in panel A. The principal steps of the pipeline are outlined in panels B, C and D. An analysis report is generated at the end of the pipeline.

Dataset validation and bootstrap resamples preparation

First, the dataset is validated and checked for sufficient number of samples (see Figure 1B). If the number of samples in a dataset is below 5 (this cutoff is an adjustable parameter) we split the dataset into 10 pseudo samples, to facilitate bootstrapping. Next, within each sample $= 1 \ldots k$, we calculate a matrix $\widetilde{M}_i$ of distances between gene pairs based on gene expression according a user-specified metric (Euclidean, cosine, or correlation distance). In a given sample, only genes that are expressed in at least 5% of cells are considered. For the choroid dataset the cutoff is 1% because of the high quality of this dataset, which was experimentally enriched for endothelial cells and had a higher sequencing depth (22). For each sample, the distance is calculated between each gene in a set of *m* genes of interest, here receptors, and each gene in the complete set of *n* expressed genes, resulting in M, an $n \times m$ matrix that contains the *n* genes in rows and *m* receptors in columns. Samples (ie, the set of *k* matrices, one per sample) are then aggregated using the median distance among non-missing values. After aggregation, the distance matrix M may still have some gene pairs with missing values, which are replaced

with 1 for correlation or cosine distances and with the maximum of M for the Euclidean distance. When the matrix M is calculated using all the available samples, we refer to this as the original data. In addition to the *original data*, we prepare a set of *bootstrap resamples* by resampling with replacement all samples N times (we used $N = 100$): i.e. we randomly choose with replacement k samples, and define a distance matrix following the same procedure as in the *original data*. The size of the bootstrap resamples was validated for one dataset, the choroid dataset, where we compared the results of 100 and 1000 resamples. The output of the bootstrap was very similar among the two independent resamples. The bootstrap frequency Pearson correlation between them was 0.99. The set of bootstrapped distance matrices is used to analyze the effect of gene expression variability and for statistical inference.

Dendrogram ordering by co-expression

The ordering of the dendrogram (see Figure 2) is based on the Pearson correlation of normalized single cell gene expression, which is a commonly used gene co-expression measure. The similarity metric is the Euclidian distance of the Pearson correlation. We also compared alternative choices for these metrics, for example we tested Spearman and Pearson correlations of gene expression (see Table S1).

Here, an $m \times m$ symmetric matrix A with no missing values is calculated from each $n \times m$ matrix M by considering the Euclidean distances of the Pearson correlations between the columns of M, thus defining a distance matrix for the m receptors only. We then perform an agglomerative clustering on the matrix A using a Ward variance minimization linkage method (23) and construct a dendrogram. (alternative clustering similarity metrics and linkage methods are compared in Table S1). Correlation is compared to the Euclidean distance of Pearson correlation in the dendrograms shown in Figures S1, S2 and S3. Finally, receptors, which are the leaves of the dendrogram, are ordered to minimize the sum of the distances between adjacent leaves in the dendrogram layout (25). This dimensionality reduction procedure defines a “dendrogram distance” between any two ordered receptors, which is a positive integer equal to the difference of their ordering index. The N bootstrapping resamples and the original data are processed according to the steps shown in Figure 1C.

Following this procedure genes that are co-expressed in individual cells are closer to each other in the dendrogram order.

The four measures shown in the lower half of the Dendrogram Figure (Figure 2) are:

1- Expression fraction

For each receptor we calculate the fraction of cells that express that receptor at non-zero level. We calculate the median across the samples. We expect important receptors to be expressed in most cells.

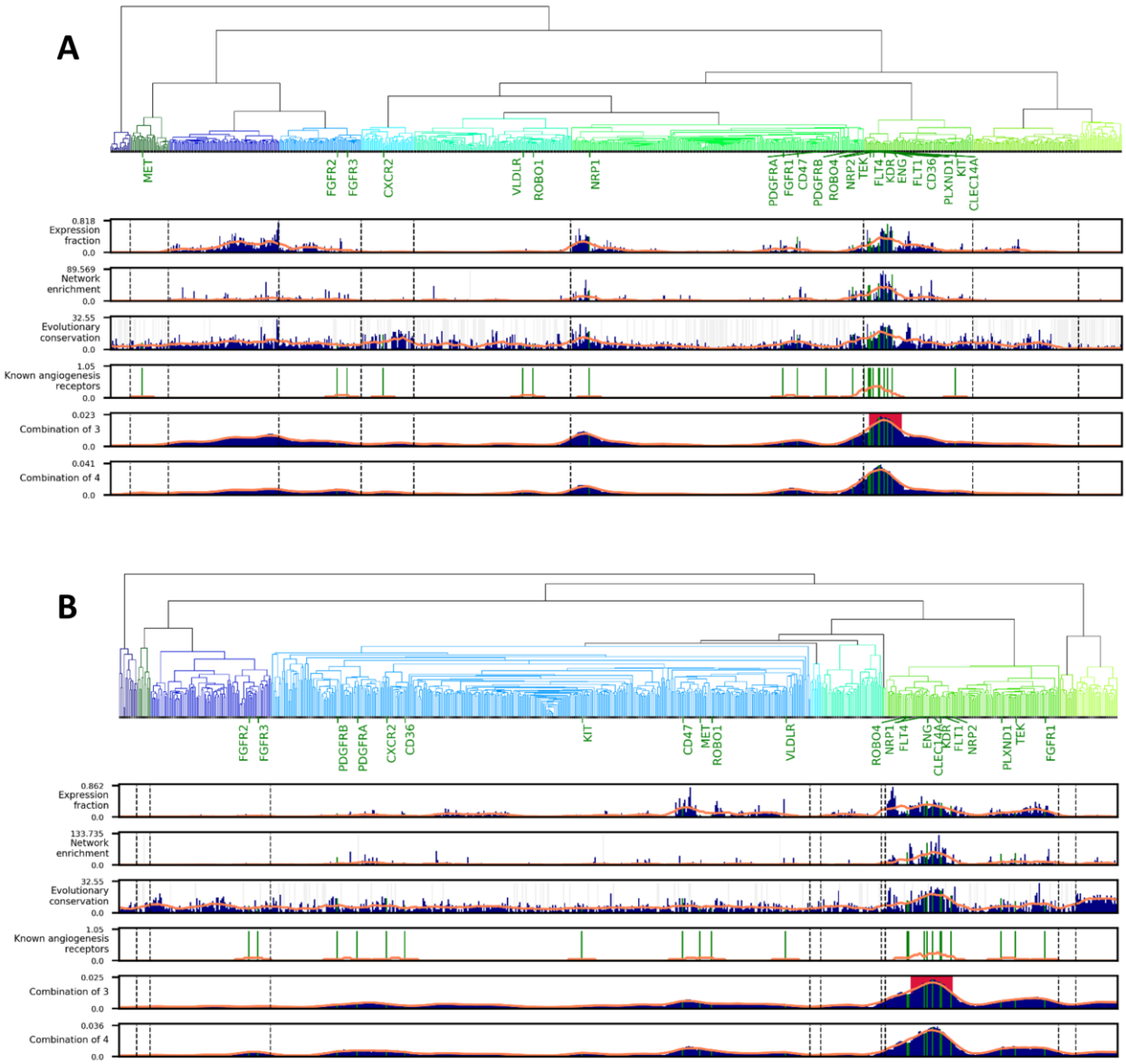

Figure 2. Top of panels A and B: ordering by single cell co-expression of receptors in endothelial cells of PanglaoDB annotated by DCS for: (A) *Mus musculus*, and (B) *Homo sapiens*. The receptors co-expression is represented by a dendrogram. The dendrogram ordering is optimized to minimize the Euclidian distance of the Pearson correlation of the receptors. Known angiogenesis regulators are labelled along the dendrogram. Bar-style panels below the dendrogram in A and B show the receptor measures and their combinations. The orange solid line in each panel is a moving window average. The main comberon is shown with a red background in the “Combination of 3” panel.

2- Network enrichment

To find receptors close to genes that are differentially expressed in endothelial cells we use a predetermined undirected gene interaction network known as *Parsimonious Composite Network* (PCN) (24). This network contains 19,783 genes and was obtained by combining data from 21 popular human gene interactions networks. Only interactions supported by a minimum of two networks were included. The 21 networks included interactions derived from multiple methods, such as protein-protein interactions, transcriptional regulatory interactions, genetic interactions, co-expression correlations and signaling relations. The authors performed an extensive evaluation of the PCN network showing that it was superior to any of the 21 networks used to build it. We independently determined that sets of genes with similar functions are close to each other within this network. Using inflammation, cell cycle and metabolic gene sets we measured intraset efficiency, a metric we have previously introduced (8) , and the enrichment for direct connections between the genes in a given set using the binomial distribution. All the tests gave highly significant p-values, supporting the high quality of the PCN network.

First, we determine the set of the top 1000 differentially expressed genes, S, in a dataset consisting of several samples. The differential gene expression for each sample is computed using a t-test between gene expression of endothelial cells and non-endothelial cell in that sample. Genes with a p-value below 0.001 are retained and ordered according to the decreasing t-test statistic. The median of the ranks across samples is taken to rank the differentially expressed genes of the entire dataset.

The t test was also compared with the Mann-Whitney U test and the network enrichment obtained was very similar (see Table S1 for bootstrap frequency obtained with the two methods).

To obtain a p-value we use the binomial distribution. Let p be the probability that a randomly chosen PCN edge is connected to a gene in the set S of differentially expressed genes. We estimate p as the number of edges with at least one gene in S divided by the total number of edges in the network. For each receptor j the probability to have k_j or more enriched edges out of its total degree n_j in PCN is given by

$$P_j(k_j) = 1 - \sum_{i=0}^{k_j-1} \binom{n_j}{i} p^i (1-p)^{n_j-i}.$$

The negative logarithm in base ten of this probability is the “Network enrichment” measure used in our method.

This measure reflects the level of expression of the first neighbors of a receptor within the PCN network. More exactly, It captures the enrichment in number of connections of a receptor to the set of differentially expressed genes.

3- Evolutionary conservation

To evaluate evolutionary conservation, we calculate the receptor dendrogram ordering using data from two different species and we quantify their similarity. The comparison is done by pairing data from different species. In this paper we use *Mus musculus* and *Homo sapiens*. For each receptor j in the first case (for example *Mus musculus*) we count the number of receptors in a certain window in the dendrogram

ordering (here of size 50, centered at j) that are also present in a window of the same size centered at homologous receptor j for the second case (for example *Homo sapiens*). The number of overlapping receptors in the neighborhoods of the receptor j defines a local measure of evolutionary conservation. Genes that do not appear in both cases are marked as missing.

4- Known angiogenesis receptors

We use for validation a set of 22 receptors that are known to regulate angiogenesis: CD36, CD47, CLEC14A, CXCR2, ENG, FGFR1, FGFR2, FGFR3, FLT1, FLT4, KDR, KIT, MET, NRP1, NRP2, PDGFRA, PDGFRB, PLXND1, ROBO1, ROBO4, TEK and VLDLR. These receptors were obtained before the analysis from a systematic Pubmed search for receptors with functional effects on angiogenesis. The list includes FLT1 and KDR, the main VEGF receptors, and also other targets for anti-angiogenic drugs under development (1, 2, 5, 25). As negative controls we used 30,836 sets of genes from the GO Biological Process class (see Table S3).

Other measures

We have considered other measures in addition to four measures listed above. These other measures include evolutionary rate and age, average gene expression, distance between neighboring leaves along the dendrogram, evolutionary conservation of known angiogenesis receptors, and known angiogenesis receptors from gene ontology angiogenesis terms. There was no clear benefit in using any of these measures in addition to the main measures. The main implementation of DECNEO allows for the inclusion of these measures, or even new, user-defined, measures in future analysis.

Combination of measures and comberons

Figure 1D shows how we analyze the combinations of different measures for expression, evolutionary conservation, network enrichment and other properties. The output is shown in Figure 2.

Two of the panels in Figure 2 show linear combinations of the measures. “Combination of 3” is a sum of “Expression fraction”, “Evolutionary conservation”, and “Network enrichment”, “Combination of 4” also includes “Known angiogenesis receptors”. The focus of this paper is on the interpretation of “Combination of 3” which is unbiased for the inclusion of specific receptors. “Combination of 4” is provided by the software for possible use in the case of smaller or lower quality datasets.

For each dataset d and for each resample c, we first normalize these measures so that the sum of each measure is one and take a moving average along the dendrogram to smooth the normalized measures (using a centered window of 21 receptors and a uniform averaging kernel). Then, we add their values and take a moving average of the combined measure to obtain a combination score m_c^d.

In the analysis, we identify the highest peak p_0 in m_c^d, such that p_0 is $\arg\max(m_c^d)$, along the dendrogram and we find all the receptors with $m_c^d \geq 0.5 m_c^d(p_0)$ that are in the immediate proximity of this main peak. Then, we count how many times a given receptor is found in the main peak across all bootstrap resamples in a dataset and we calculate its frequency of peak inclusion. The software can also calculate the frequency of inclusion for the main and all other secondary peaks, p_i where $i \geq 0$, in the dendrogram by

considering all extrema of height and prominence larger than 0.05 separated by at least 50 receptors. Prominence of a peak is defined as the vertical distance between the peak and the lowest contour line encircling it but containing no other higher peaks. When secondary peaks are used, each peak i (main or secondary) contains at most 50 receptors that are in the immediate proximity of the peak and have $m_c^d \geq 0.5 m_c^d(p_i)$. See Figures 2 and S4.

The sets of genes that are included in the peaks are the comberons. It is important to observe that the measures used to identify the peaks are not necessarily increased for a particular gene. For example, as can be seen from Figure 2, a gene that has a low value of network enrichment can be included in a peak because is co-expressed with other genes that have a high value for this measure.

Visualization

A "plot" mode (see Figure 1C and in part Figure 2) can be selected to visualize dendrograms, heatmaps of the A matrix (as an option), and several other panels. These panels plot the measures described above, such as the Evolutionary Conservation and Expression Fraction, smoothed using a moving average along the dendrogram (indicated by an orange line), the combinations of these measures and highlight the location in the dendrogram of the known angiogenesis receptors.

Statistical analysis of Dendrogram Ordering

The likelihood-ratio test was used to analyze the distance from KDR and FLT1 of known angiogenesis receptors in the dendrogram. This test compares the likelihood of outcome under the null hypothesis, $L(N)$, with the likelihood of outcome under the alternative hypothesis, $L(A)$. The null hypothesis here is that the known angiogenesis receptors are uniformly spread in the optimally ordered dendrogram, whereas the alternative hypothesis is that the known angiogenesis receptors should be in the vicinity of KDR (or FLT1) in the dendrogram, showing that known angiogenesis receptors are co-expressed. Under the null hypothesis every dendrogram receptor has equal probability to be one of the known angiogenesis receptors. With the alternative hypothesis we use a logistic regression to get the probabilities of the receptors to be a known angiogenesis receptor given the optimal dendrogram ordering. The statistic of likelihood-ratio test is $-2\ln(L(N)/L(A))$, and the p-value is calculated from the survival function of chi-squared continuous random variable.

An alternative non-parametric method was also used for the same purpose (26). The robust Poisson pseudo maximum likelihood estimation (PPMLE) was implemented in R using the package *gravity* (27) . The p-value was obtained from the deviances of the null hypothesis and alternative hypothesis using the difference test in the chi square distribution.

I

Statistical analysis of Comberons

The main aim of our statistical analysis was to show that there is a set of receptors significantly associated with the largest peak of Figure 2. We call the set of receptors associated with the largest peak a comberon.

The first step was the use of the bootstrap method (28) to improve the robustness of the dendrogram. This is a very common procedure in many scientific fields where

dendrograms are used (29). In fact the paper describing the application of the bootstrap method to dendrogram analysis is one of the 100 most highly cited papers of all time (30, 31)..

The bootstrapping involved 100 resamples with replacement. The values shown in Table 1 are bootstrap frequencies where 1 means that the receptor was found in the largest comberon in all of the 100 resamples. See Table S4 for the complete list of the bootstrap frequencies.

We then performed several statistical tests:

1- A permutation (also called randomization) test (32, 33) for the co-expression that was used to build the dendrogram. We wanted to determine the statistical significance of the frequencies in the first column of Table 1, which contains the *Mus musculus* data analyzed with DCS, and we therefore used a control where the ordering of the dendrogram was no longer dependent on correlation of gene expression of the receptors. First, we determined the maximum frequency of genes included in the main peak for 100 instances (to match the bootstrapping procedure) with randomized ordering of the dendrogram shown in Figure 2A. We then repeated this procedure for 10^7 realizations. This calculation was done using the High Performance Computing Center (HPCC) supercomputer at Michigan State University. From the 10^7 realizations we obtained a distribution of the maxima of the bootstrap frequencies for the original data case. We did not observe a frequency above 0.5 in any of the 10^7 realizations. The maximum observed frequency (0.49) was slightly below this number as shown in Figure S5. This procedure allowed us to obtain a non-parametric estimate of the probability of obtaining a frequency above 0.5. Note that this is a conservative estimate of the p-value given, because it is an estimate for a single value being larger than 0.5 while in first column of Table 1 there are 18 genes above this threshold. We also empirically determined that this distribution resembles a generalized extreme value distribution, which is a known distribution of maxima (see Figure S5).

2- As a validation we split the dataset in two equal parts. This was done twice, once splitting the samples and once splitting the studies. The bootstrap frequency of the two splits were then compared.

3- The significance of enrichment of the comberon for known angiogenesis receptors was calculated using the hypergeometric test. This was done for both the original and the bootstrap frequency of at least 0.5.

4- We also obtained an independent dataset composed of 14 single cell mice gene expression studies, containing 71 samples and 65,057 endothelial cells (see Table S5). The endothelial cells were identified using DCS as was done for the original Panglao dataset. This independent dataset was obtained after the analysis of the main dataset had already been completed and did not affect any of the methodological choices. The evolutionary conservation was obtained comparing the independent mice data with the human data from PanglaoDB. The bootstrap frequency of the independent dataset was compared with that of the original dataset.

Implementation of the algorithm

The pipeline shown in Figure 1 is implemented in our new software DECNEO (DEndrogram of Co-expression with Evolutionary Conservation and Network Enrichment Ordering). We have developed a main and an alternative implementation of DECNEO that are open-source and maintained at https://github.com/sdomanskyi/decneo. Releases are published on the Zenodo archive.
The main implementation can be installed from PyPI using the command "pip install decneo". DECNEO modules are relying on Python numpy, scipy, pandas, scikit-learn, matplotlib and other packages (see Figure S6 for dependencies). The main implementation can leverage on a multicore large-memory High-Performance Computing Systems for fast processing of large datasets. The detailed documentation and installation troubleshooting instructions are available at https://decneo.rtfd.io/. The processed data and results reported in this paper are deposited at https://figshare.com/. The source code of the alternative implementation is deposited at https://github.com/sdomanskyi/decneo/validation. The software was optimized for use with the HPCC supercomputer at Michigan State University but also tested on different operating systems and hardware.

Alternative implementation

The main implementation of DECNEO was validated by comparing its results to an additional implementation created independently as a software and methodology validation. The alternative implementation has allowed us to verify the robustness of the results under minor differences in the software implementation. Specifically, the differences are: (1) When finding differentially expressed genes for the network enrichment, the alternative implementation filters out genes that are downregulated and sorts them by p-value while the main implementation sorts genes by the median rank of the t-test statistic. (2) When computing the expression fraction for a set of samples, the median of the expression fraction across samples is taken, instead of the mean across samples as in the main implementation. (3) Lastly, in the alternative implementation, gene measures are not smoothed before merging them into a combination of measures, and the combination of measures is normalized so that the minimum is zero and the maximum is one. (4) In the alternative implementation only endothelial cells, but not non-endothelial cells, are used when generating pseudo samples.

The analysis of endothelial cells in PanglaoDB-Alona dataset is summarized in Figure S7 and the alternative implementation is shown in Figure S8. These analyses show results similar to Figure 2. Table S6 shows the correlation of bootstrap frequencies between DECNEO main and DECNEO alternative implementation.

Revealing Multiple Peaks

We observe multiple peaks in the measures and their combinations. See for example Figure S4, which corresponds to "Combination of 3" of *Mus musculus* endothelial cells of PanglaoDB-DCS. We analyzed all the peaks, including the largest one. Each instance of the 100 bootstrap resamples has a set of peaks with the receptors identified in the same way as in Figure S4. We order receptors (vertical axis) and peaks (horizontal axis) according to the "Ward" agglomeration method with Euclidean metric of binary membership, see Figure S9, which is colored according to each peak relative height. The left and top dendrograms highlight receptors and peaks groups. To focus on the most

evident peaks we obtained 15 groups of receptors and 15 groups of peaks. After determining the density of the receptors in each receptor-peak group we show in Table S7 only those with at least 0.5 density. All the groups of peaks that correspond to one group of receptors are aggregated to determine average frequency and average peak height.

Tissue-specific receptors evaluation

We have used the analysis pipeline, illustrated in Figure 1, separately on several tissue-specific studies (SRAs) from PanglaoDB annotated by DCS, from *Mus musculus* and *Homo sapiens* datasets. In order to compare the processed SRAs, we calculated the Pearson correlation of the bootstrap frequencies, shown as a heatmap in Figure 3. The SRAs are ordered according to the "Ward" (23) agglomerative clustering of the Euclidean similarity metric of the Pearson correlation of bootstrap frequencies.

The top tissue-specific receptors are shown as a heatmap in Figure 4, to show that some receptors can be found in more than one tissue but other are more specific. SRA studies of the same tissue were combined by averaging the largest peak bootstrap frequencies. The lists of the top-10 receptors for each of the analyzed tissues and the bootstrap frequency cutoffs are given in Table S8.

Metascape enrichment analysis

For pathway enrichment analysis, Metascape (34) calculates the p-value defined as the probability from the cumulative hypergeometric distribution and corrects it with the Bonferroni correction factor to give the q-value.

RESULTS

Overview of single cell gene expression data

We used the DCS pipeline (20, 21) to process the single-cell RNA-seq PanglaoDB database (18). The pipeline was applied to each sample independently, without correcting for batch effects, to normalize the data and identify endothelial and non-endothelial cells. Figure 5A shows 101,470 *Mus musculus* endothelial cells from 397 samples within 51 studies projected on a t-SNE layout color-coded by tissue of origin, following the Panglao tissue labels. The proportion of endothelial cells represented by each tissue type is shown in the bar plots. A large fraction of the cells in this dataset is from brain tissues. Figure 5B summarizes the *Homo sapiens* portion of the PanglaoDB endothelial cells identified by DCS. There are much more data from *Mus musculus* studies than *Homo sapiens* resulting in higher statistical power in the analysis of *Mus musculus* tissues. In the control and validation analysis we also use the PanglaoDB original annotation of endothelial cell by the Alona software. The Alona analysis of the PanglaoDB database results in a different number of samples containing endothelial cells. We will refer to the PanglaoDB data processed using DCS and Alona as "PanglaoDB-DCS" and "PanglaoDB-Alona", respectively.

In the tissue-specific analysis we use the single cells dataset from Voigt et al. (22) containing *Homo sapiens* choroid data from 18,621 cells in 8 samples. The samples were enriched for endothelial cells, mapped by CellRanger and then normalized and annotated

by DCS. Figure S10A-B shows a t-SNE layout and the relative proportions of choroid cells colored by cell type. Although this dataset contains 8 samples, there are no strong batch effects, and the main group of cells, mainly endothelial, includes all 8 biological samples, Figure S10C.

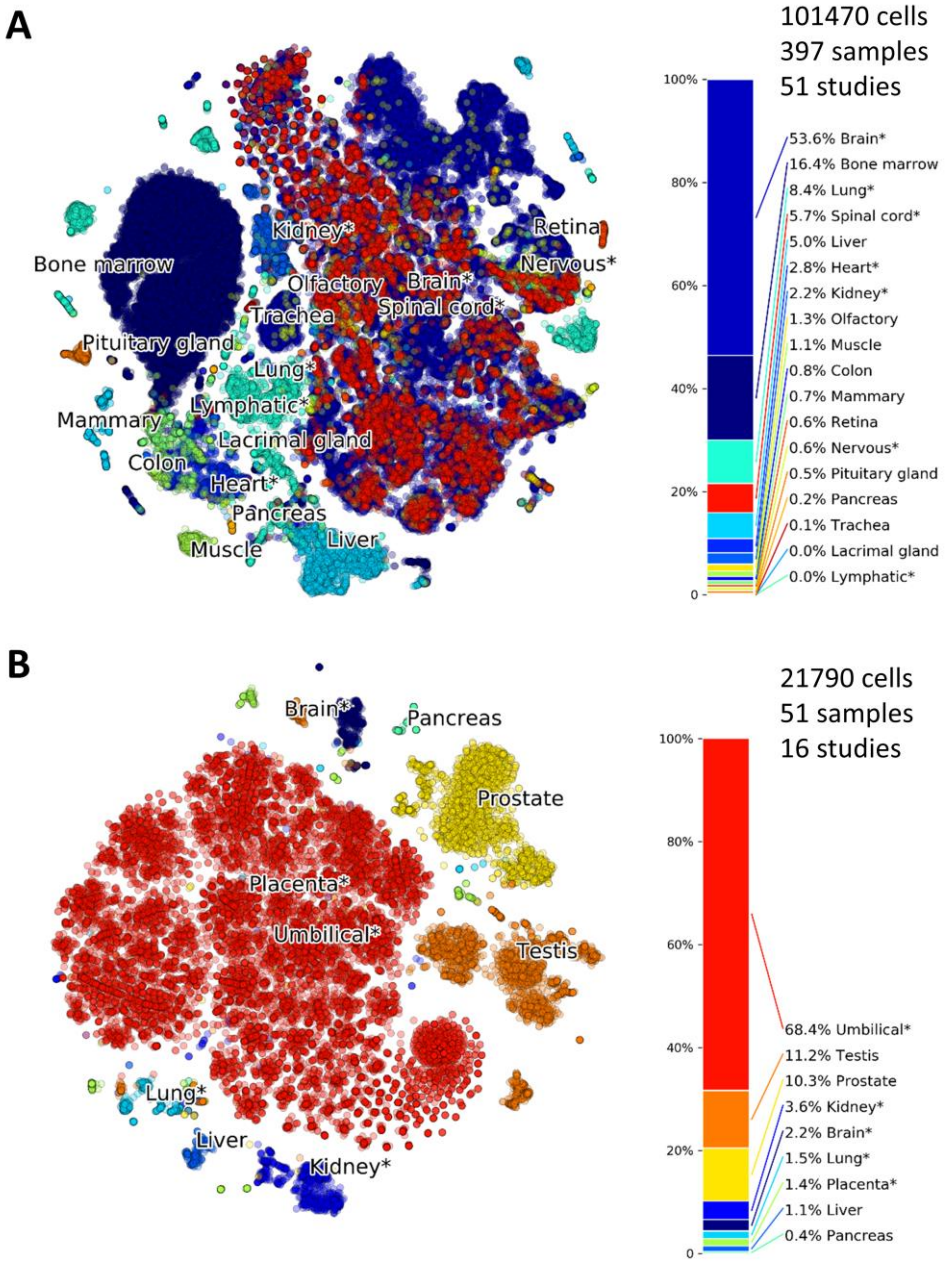

Figure 5. 123,260 endothelial cells identified with DCS from PanglaoDB that collects 22,382,464 cells in total of 1,368 samples, in 236 studies. Panels show t-SNE projection and bar chart of the relative fractions of the endothelial cells color-coded by tissue of origin, separately for: (A) *Mus musculus*, and (B) *Homo sapiens*. The asterisk after the tissue names is to denote that the PanglaoDB data is includes multiple sub-tissues that are here combined into one. The DCS processing pipeline was used for each of the 1,368 samples independently.

Dendrogram of co-expression ordering

The optimal receptors ordering is an important step in our algorithm, since we aim to determine groups of co-expressed receptors.

The subset of the "Known angiogenesis receptors" that are expressed in endothelial cells in the PanglaoDB-DCS *Mus musculus* datasets is shown in a panel below the dendrogram in Figure 2A. These known angiogenesis regulators localize in the ordered dendrogram near each other and near KDR and FLT1, the two main VEGF receptors. The p-value (Likelihood-Ratio test) of a logistic regression of the known angiogenesis receptors distance from KDR and FLT1 as compared to a uniform distribution is less than $7 \cdot 10^{-4}$ and the AUC of ROC is 0.74 for both KDR and FLT1, indicating that the dendrogram ordering in part captures the fact that these receptors participate together in angiogenesis.

The non-parametric robust PPMLE was also applied to test the distance of the known angiogenesis receptors from KDR and FLT1 in the dendrogram. Both tests gave a p-value < $5.4 \cdot 10^{-6}$.

Genes in the main peaks (comberons) and bootstrap analysis

Using the "Combination of 3" measure of the PanglaoDB-DCS *Mus musculus* data we determine which of the 898 expressed receptors are localized in the largest peak identified by this quantity. Receptors that belong to the largest peak are highlighted in blue in Figure S4, and include: ADGRF5, ADGRL4, ADRB2, AQP1, CALCRL, CD300LG, CD36, CD40, CD74, CD93, CDH5, CLEC14A, CLEC1A, CXCL16, ENG, ESAM, FLT1, FLT4, GPIHBP1, GPR20, IL4R, ITGA4, KDR, LYVE1, NPR1, NRP2, OSMR, PECAM1, PLXND1, PTPN18, PTPRB, RAMP2, ROBO4, SIRPB1, STAB1, TEK, TIE1 and TWF2. These 38 genes include 10 of the 22 known angiogenesis regulators (including KDR and FLT1, the two main VEGF receptors), and the probability of such an outcome by chance is less than $2.4 \cdot 10^{-9}$ by hypergeometric test. This is an indication that the largest peak captures receptors that are important in angiogenesis. The statistical significance is higher than that for the distance from KDR and FLT1 calculated in the previous section, showing that the comberon is a useful concept for functional elucidation of receptor expression.

To quantify the significance of a receptor being included in the main peak (the comberon) in the ordered dendrogram we perform a permutation analysis of the coexpression values after the boostrap. See Table 1 for the genes with the largest bootstrap frequency. A hypergeometric test similar to that done above on the comberon from the original dataset was also done on the top genes from the bootstrap shown in Table 1 in the first column. In this set there are 7 known angiogenesis receptors and this gives a p value of $5 \cdot 10^{-8}$. As negative controls we used 30,836 sets of genes from the GO Biological Process class (see Table S3). This table shows that the only strongly significant associations are with terms related to angiogenesis.

The bootstrap frequencies of *Mus musculus* and *Homo Sapiens* PanglaoDB annotated by DCS and Alona and with the alternative DECNEO implementation using Alona are correlated, see Table 1 and Table S6, showing consistency among the methods. Table 1 and Table S4 show that the highest frequencies are obtained in the *Mus musculus*, as expected given the larger number of datasets.

Table 1. Main comberon bootstrap frequencies of PanglaoDB endothelial cells annotated by DCS and Alona, *Mus musculus* and *Homo Sapiens*, combination of 3 and 4 measures. The table is sorted by the first column in descending order. Only receptors where the first column values are above 0.5 are shown (this corresponds to p-value of less than 10^{-7}). For complete list see Table S4.

	DCS				Alona				Alona (alternative implementation)			
	Mus musculus		*Homo sapiens*		*Mus musculus*		*Homo sapiens*		*Mus musculus*		*Homo sapiens*	
	3	4	3	4	3	4	3	4	3	4	3	4
RAMP2	1	1	1	0.99	0.99	0.97	0.41	0.3	0.53	0.51	0.41	0.27
ESAM	1	1	1	0.99	0.97	0.94	0.55	0.4	0.55	0.55	0.59	0.45
PECAM1	1	1	1	0.99	0.99	0.99	0.89	0.7	0.96	0.96	0.85	0.77
ADGRF5	1	1	0.98	0.97	1	1	0.89	0.69	0.92	0.94	0.59	0.73
CDH5	1	1	0.91	0.88	1	1	0.44	0.37	0.95	0.95	0.51	0.67
CD93	1	1	0.84	0.82	1	1	0.2	0.22	0.95	0.95	0.83	0.86
KDR	1	1	0.71	0.76	0.95	0.96	0.33	0.38	0.94	0.95	0.56	0.7
ENG	1	1	0.91	0.93	0.97	0.98	0.48	0.61	0.96	0.96	0.65	0.8
FLT1	1	1	0.95	0.96	1	1	0.55	0.64	0.87	0.9	0.58	0.73
ADGRL4	0.99	0.99	0.98	0.97	1	1	0.73	0.68	0.96	0.96	0.87	0.77
PTPRB	0.97	0.98	0.98	0.97	1	1	0.57	0.59	0.89	0.93	0.72	0.8
TEK	0.94	0.94	0.05	0.1	0.73	0.77	0.16	0.27	0.91	0.9	0.03	0.08
TIE1	0.94	0.94	0.98	0.98	0.53	0.5	0.34	0.33	0.08	0.06	0.42	0.52
CLEC14A	0.75	0.77	1	0.99	0.48	0.52	0.74	0.66	0	0	0	0
ROBO4	0.75	0.78	0.45	0.54	0.49	0.48	0.51	0.61	0.11	0.11	0.13	0.22
GPIHBP1	0.62	0.65	0.21	0.18	0.19	0.27	0.05	0.05	0.01	0.02	0.27	0.23
FLT4	0.57	0.6	0.1	0.16	0.12	0.14	0.52	0.59	0.07	0.16	0.04	0.08
CALCRL	0.53	0.57	0.89	0.87	0.51	0.49	0.28	0.27	0.08	0.07	0.12	0.17

Receptors that have *Mus musculus* PanglaoDB-DCS “Combination of 3” bootstrap frequencies of at least 0.5 are included in Table 1. The complete frequencies data are available in Table S4. A permutation test (see Statistical analysis section of Methods) allowed us to obtain a non-parametric estimate of the probability of obtaining a frequency above 0.5 as $p< 10^{-7}$ for column 1 of Table 1. A lower probability set ($p<0.05$) can also be identified with bootstrap frequency values between 0.33 and 0.5. This includes an additional 9 genes shown in column 1 of Table S4.

Validation of the comberon

Validation was obtained by splitting the original Panglao dataset in two equal parts and by obtaining an independent dataset.

The results from the split datasets are shown in Table S2. The split was done in two ways, one where the splits had the same number of studies and the other where they had the same number of samples.

For the splits containing the same number of studies the Pearson correlation of bootstrap frequencies between the two splits was 0.87. The number of genes with bootstrap frequency above 0.5 in the first split dataset that was also found in the second split was 17 out of 23. This gives a p value of $1.5 \cdot 10^{-27}$ using the hypergeometric test.

For the splits containing the same number of samples the Pearson correlation of bootstrap frequencies between the two splits was 0.91. The number of genes with bootstrap frequency above 0.5 in the first split dataset that was also found in the second split was 16 out of 26. This gives a p value of $3.3 \cdot 10^{-24}$ using the hypergeometric test.

For the independent dataset we obtained 14 additional single cell mice gene expression studies, containing 71 samples and 65,057 endothelial cells (see Table S5). The bootstrap frequencies of the independent dataset were compared to those of the original dataset (first column of Table1). The comparison is shown in Table S12. The Pearson correlation of bootstrap frequencies between the original and the independent datasets was 0.85. The number of genes with bootstrap frequency above 0.5 in the main dataset that was also found in the independent dataset was 16 out of 25. This gives a p value of $1.6 \cdot 10^{-27}$ using the hypergeometric test. The dendrogram for the independent dataset is shown in Figure S12.

Comberons from Multiple Peaks

In addition to the main peak analyzed in Table 1 multiple other peaks can also be detected. Using the “Combination of 3” of *Mus musculus* endothelial cells of PanglaoDB-DCS we identified the Group with the highest bootstrap frequency and peak height. This Group (Group P1 in Table S7, blue cluster in Figure S9) is composed of the same 13 receptors which had the highest bootstrap frequency in Table 1, column 1. In addition to this main group of receptors we identified 10 other receptor groups (comberons), shown in Table S7. The names of the groups, P1 through P11, reflect ranking in their average bootstrap frequency.

We performed Metascape analysis on the 11 groups and found that groups P1, P6, P9 and P11 are all enriched for blood vessel development, cell migration and growth, and peptidyl-tyrosine phosphorylation, see Figure S11. The detailed KEGG, GO and Reactome gene enrichment analysis of the 11 groups, Table S9, using PyIOmica (35) enrichment functions confirms the results of the Metascape analysis. The evidence for the biological functions of the minor comberons is however not as strong as for the main comberon.

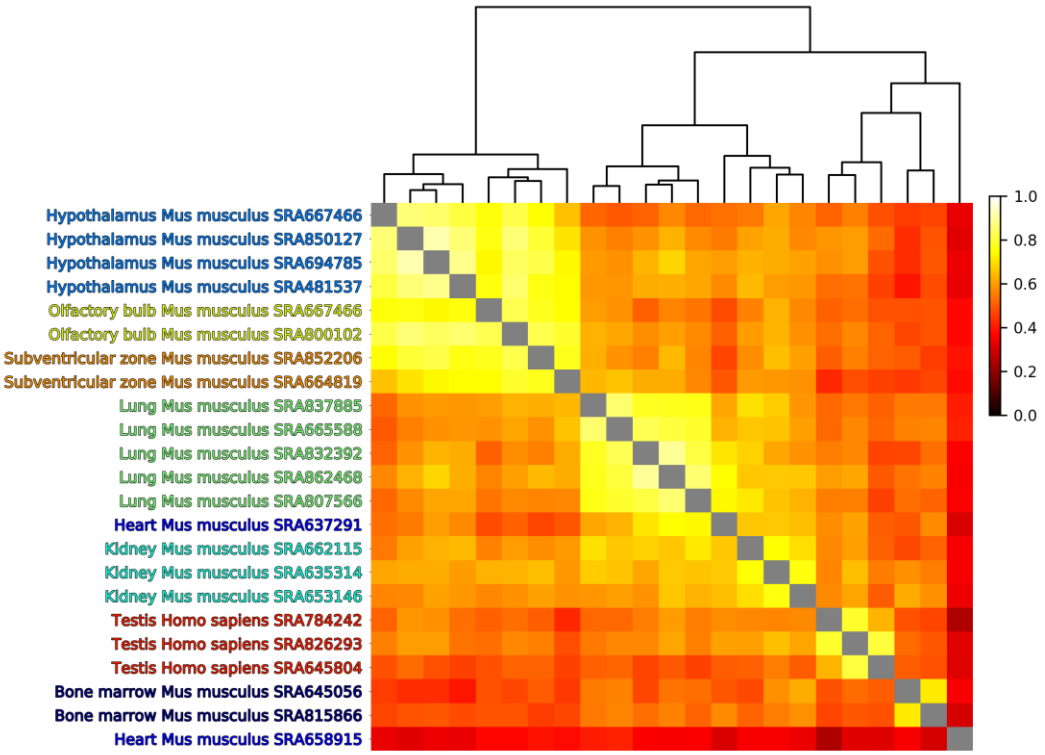

Figure 3. Analysis of endothelial cells from selected studies in PanglaoDB database annotated by DCS. Clustering of Pearson correlation of the largest peak bootstrap frequencies for the combination of 3 measures. Studies from most similar tissues group together.

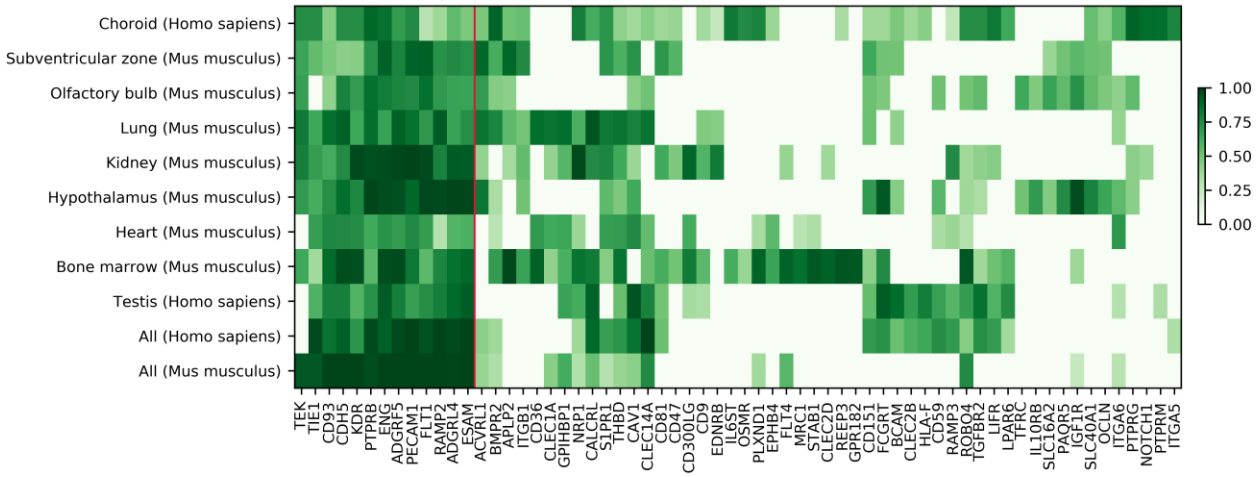

Figure 4. Comberons of endothelial cells of specific tissues from PanglaoDB and choroid annotated by DCS. Under All we include the global analysis shown in Table 1 for comparison. The 13 receptors to the left of red line are the 13 top receptors from Table 1. The rest of the receptors are within the top-10 tissue-specific genes for at least one set.

Tissue-specific receptors

There are 8 *Mus musculus* and *Homo sapiens* tissues within the PanglaoDB database annotated by DCS for which there are at least two studies with at least 300 endothelial cells each including at least 5 samples. The correlation among the 21 studies within these 8 tissues is shown in Figure 3. Most studies belonging to the same tissue cluster together. Additionally, the hypothalamus, olfactory bulb and subventricular zone studies, which are all parts of the brain, also cluster together. The average correlation values within tissues of the same type are almost always larger than the average correlation with other tissues (see Table S10). These results indicate that endothelial cells have tissue specific co-expressed receptors.

Figure 4 shows the receptors with the highest bootstrap frequencies (the comberons) for the 7 tissues shown in Figure 3 and for one addition human dataset obtained from the choroid of the eye (22). These tissues usually also have high bootstrap frequencies for the top 13 receptors from the general analysis of Table 1 (see receptors left of the red line in Figure 4).

Two tissues, lung and choroid, were selected for further analysis using Metascape. To focus on tissue specific pathways, we did not include the generally relevant 13 genes shown on the left of Figure 4. In the lung we found the “cell-matrix adhesion” GO term as the top significantly enriched term (Table S11), with q-value of less than 10^{-6}. In the choroid, Metascape enrichment analysis yielded blood vessel morphogenesis as a top significant GO term (Table S11), with q-value of less than 10^{-9}. By removing the receptors contained in the top enrichment term we found that the second most significant GO term for lung is cyclic-nucleotide-mediated signaling and for choroid is the JAK-STAT signaling pathway, (Figure S13).

DISCUSSION

Summary

We can now address the questions presented at the end of the Introduction. Our results show that single cell data can indeed help us to identify a set of co-expressed endothelial cell receptors, conserved among species (mice and human) and enriched, within a network, of connections to up-regulated genes. This set does include VEGF receptors and includes several other receptors previously shown to play a role in angiogenesis. We can combine endothelial single-cell data from multiple organs and show organ-specific receptor profiles in these sets.

When referring to the sets of co-expressed receptors identified by our approach, we found useful to coin a precise term. We suggested the term receptor comberon, by analogy with the term operon introduced by Jacob and Monod for a different type of co-expressed genes, coded by the same mRNA in bacteria (10) so that we could highlight our intellectual debt to their pioneering work.

Many advances have contributed to increase our understanding of the molecular machinery coordinating gene expression (36, 37), but we know less about the functional effects of combinations of expressed genes in the human body. Our method identifies combinations of receptors integrating co-expression in single cell with 3 additional measures: expression fraction, network enrichment and evolutionary conservation.

Importantly, these 3 measures do not need to be increased in all members of a comberon, so that the method is able to deal with incomplete information and with evolutionary change. The procedure has led to sets including clinically validated drug targets, like the VEGF receptors, pointing to the key strategy for the full elucidation of the functional significance of the comberons.

The conclusions are supported by multiple highly significant statistical tests obtained from large datasets, including many independent experiments. Statistical validation was reinforced by splitting the original dataset in two parts, each analyzed separately and by obtaining an additional independent set of single cell gene expression data. Evolutionary pressure to preserve functionally important and biologically robust combined effects provides a plausible mechanism for the evolutionary conservation of the main endothelial receptor comberon between mice and humans. The identification of receptor combinations is also consistent with the need for anti-angiogenic drug combinations emerging from clinical experience.

We have also adopted a strategy of extensive internal replication to further validate our findings. Two versions of the software implementation of the algorithms have been written independently by two of the authors (SD and AH), to guard against coding errors and unforeseen consequences of implementation choices. We have also used two different methods to identify endothelial cells in single cell RNAseq data: DCS, (20, 21) which was previously developed by some of us, and Alona (19). Other validation strategies also represent replications in a broader sense. To identify evolutionary conservation, we studied similarities in comberons among two species, mouse and humans. We also found in specific tissues many of the receptors that were present in the global comberons. The unbiased strategy to identify comberons (called Combination of 3 in Figure 2) showed a strongly significant enrichment for the presence of receptors already known to play a role in angiogenesis. Finally, the Discussion was written by and for scientists from multiple disciplines, providing diverse intellectual perspectives supporting our conclusions.

Biological implications

In many scientific disciplines it is normal to examine the results of an experiment in the light of general models. For our biological results we can to refer to a recent description by Paul Nurse of the five major ideas that underpin Biology (13). Paul Nurse received the Nobel Prize for his cell biology discoveries, and he is the Director of the Francis Crick Institute. He has stated that the ideas he has presented are not new but that even for an experienced scientist like him putting them together was illuminating (13). Our work has interdisciplinary implications and it is especially helpful for non-specialists to see the results from a wider perspective.

The five fundamental ideas of Biology according to Paul Nurse are the cell as the basic unit of life, the gene as the basis of heredity, evolution by natural selection, chemical activities bringing about life and life as managing of information. The approach described here touches on all of them, including single cell measurements of gene expression, evolutionary conservation, molecular interactions of receptors and biological network models of information transmission and integration. It is therefore not surprising that the approach would eventually prove informative, but it is still important to show that available datasets can already provide a clear signal.

If a set of co-expressed receptors is evolutionary conserved it is reasonable to ask why. The main set we have identified (see Table 1) includes VEGF receptors and several other receptors clearly involved in angiogenesis. In mice, where more data are available, it includes at least 18 receptors. This redundancy is common in biology and has been shown to lead to robust control (9, 38, 39). It might also be byproduct of the tinkering action of evolution, which often proceeds by gene duplications and can lead to not fully adaptive solutions (40, 41). A classic example is the blind spot of the human eye, which is consequence of the inverted retina, not present in the octopus (42).

Many of the receptors in this set are known to promote angiogenesis however some (like PLXND1 in the lower probability set (43, 44)) are considered to be angiogenesis inhibitors. This is not surprising if we consider that biological regulation aims to obtain homeostasis and controlled response, rather than uncontrolled activity (13). In fact, it has been already reported that MEF2C, a transcription factor induced by VEGF, inhibits angiogenesis, providing negative feedback (45).

Pharmacological and medical implications

As we have mentioned in the Introduction, the problem of finding anti-angiogenic drug combinations is still largely unsolved, as it is also for many other complex diseases and conditions. Traditionally drug targets are individual proteins but mimicking the physiological control of cells would suggest that a comberon could be the appropriate drug target for complex diseases. It is not unfeasible to target receptor comberons, because Ab drug inhibitors for each receptor can be developed. There are around 1,000 human cell membrane receptors (46) (excluding olfactory receptors and other specialized types) and until now Ab drugs have been approved by the FDA against 26 of them (47). Our work supports the substantial acceleration of Ab drug development against receptors and their ligands, because it suggests a new way to use these drugs, initially as research tool and later as therapies. While in vitro models, for example organoid models using iPS cells, can be informative, they can only be stepping-stones to clinical studies if we want to elucidate human intercellular communication *in vivo*. The last step can only be achieved with therapeutic-grade antibodies.

This is clearly only one more reason for Ab drug development acceleration, in addition to several others, including the increasing role of Ab as individual drugs (48) and the prospect of advances in decoding of intracellular signaling, possibly using methods similar to those used to decipher encrypted messages (49).

The process used by Shinya Yamanaka and his group to identify combinations of factors to induce pluripotent stem cells can be used as an informative analogy for therapeutic interventions on comberons (15). Starting from a longer list of transcription factors he used several methods, including gene expression, to identify a set of 24 transcription factors that he then reduced experimentally, finding that just 4 of them were sufficient (15). Even if receptor comberons are quite large, biological systems respond dynamically and acting on just some of the receptors might be sufficient to induce a stable change from pro to anti-angiogenic activity. It is likely that this last step will need to be conducted experimentally as was the case for the Yamanaka factors.

Beyond angiogenesis, it is clear that important cellular processes are mediated by a large number of stimulatory and inhibitory signals mediated by membrane receptors. Different cell types form complex and in part organ-specific intercellular signaling networks. For

example, the response to injury and infection is mediated not only by the cells of innate and adaptive immunity but also by other interacting cell types like endothelial cells, epithelial cells and fibroblasts (50, 51). Our understanding of these interactions is however incomplete, and our therapeutic tools are limited and do not match the complexity of natural signals.

We are writing this while the first promising results from the phase 3 COVID-19 vaccine trials are starting to appear and thoughts are starting to turn to the broader issue of pandemic preparedness. It has been noted that the past decade has witnessed unprecedented pandemic explosions and that we have entered a new pandemic era, due to a complex set of causes including human population growth, increased mobility and environmental factors (52). The COVID-19 pandemic has stimulated our understanding of the importance of controlling the response of endothelial and immune cells to infectious agents. For example, Ackerman et al (53) have recently shown that COVID-19 patients exhibit severe endothelial injury and perivascular T-cell infiltration. An imbalanced host immune and inflammatory response drives the development of COVID-19 (54). The randomized RECOVERY trial (55), showing the beneficial effect of dexamethasone on COVID-19 mortality, provides perhaps the strongest evidence about the clinical advantage of strategies acting on the immune response, which we cannot however yet calibrate precisely. Developing pharmacological tools to control the signals among cells of the immune and inflammatory system is an aspect of pandemic preparedness that might be deployed as part of a fast response, before virus-specific drugs and vaccines are available.

Social and economic implications

While developing Ab drug is feasible, a clear barrier for a biomimetic strategy targeting receptor comberons is the very high and increasing cost of drug development (56). Combining drug development with fundamental scientific understanding of *in vivo* human intercellular communication is challenging. As a research area that combines elements of basic platform science and its application to therapeutics for specific condition, it does not fit neatly into the distinct priorities of the two largest sources of US biomedical funding, federal support (mainly NIH) and industry.

A recent report from the National Academy of Science (57) pointed out that the benefits of some innovative projects have previously been accepted slowly due to resource limitations. They state that “much of the biomedical research community was strongly opposed to the Human Genome Project when it was first proposed, believing that it diverted resources from more valuable investigator-driven work (58). The project and its impact look much different in hindsight.” (57). Diverting public resources from valuable scientific programs is however a legitimate concern. The academic community is also not equipped to take drugs all the way to approved clinical products and is not likely to be supported by NIH to do so.

Industry funding might also be only partially appropriate. The fundamental nature and the complexity of the problem of intercellular communication within organisms, including human beings, would benefit from full scientific openness and sharing (57). This complete openness is not possible in the pharmaceutical industry especially at the preclinical stage. Even recent high-profile papers from pharmaceutical scientists showing that many academic studies are not reproducible were not able to specify the studies they

were referring to (59, 60). This is an obstacle to the self-correcting mechanisms of science. Evidence that Celera's gene-level intellectual property hindered subsequent research investments and innovation offers a cautionary tale in this regard (61). The case of genetically-modified mice also suggests that openness can increase innovation through greater exploration of novel research directions (62). It is precisely this type of exploration that is more likely to yield breakthrough innovations (63). The key to ensuring a platform that can enable a wide range of scientific discoveries and therapeutic advances is to place it in the public domain such that would be innovators can access that knowledge with few barriers. The platform will be composed of approved Ab drugs for most receptors and, giving the increasing cost of individual drugs (56), using only industry resources is likely to make the cost of large combinations unsustainable.

The best opportunity is provided by large philanthropic sources of funding, which are able to combine non-profit motives with industrial collaborations. We refer specifically to the increasing number of billionaires joining the Giving Pledge (64). These philanthropists (211 at present) have committed to donate the majority of their wealth. According to their statements (64) they are looking for unique projects with a potential large impact on society that are not supported by other sources and many also wish to contribute their entrepreneurial skills. The Giving Pledge was started in 2010, however a recent report of the Institute for Policy Studies (65) has found that the total wealth of most of the Pledgers has increased since 2010. Some have however provided large donations and there is also an example, that of Charles Feeney, who has donated his entire wealth (66). Even using a conservative estimate the total of available funds is larger than ten times the annual budget of the NIH, the biggest supporter of academic biomedical research in the US (65). The shortage of suitable projects with a potential large impact that Pledgers would support could be addressed by one or more joint initiatives of the biomedical community.

Cell communication is a fundamental unsolved biomedical problem and Ab drugs could be provided as a platform technology to academic and industrial groups, resetting the traditional border between basic science and drug development. Providing tools rather than directing research is consistent with many previous examples of scientific innovation suggesting that a diversity of approaches should be encouraged (63). The efforts of other scientific communities that have been able to define priorities for large-scale projects might be informative. An example is the Snowmass process (67) which has been used by particle physicists for several decades and has guided the choice of major projects after encouraging discussions and feedback from all scientists with relevant expertise. The Snowmass process aims to produce a scientific consensus, as encouraged by federal funding agencies, especially the DOE (67). An alternative that was able to facilitate and record the expression of a large diversity of opinions in addition to the preferred choices, using dedicated technology, is provided by a recent survey of more than 4,000 biomedical scientist (68). The scientific community could greatly promote human health and the efficient use of large philanthropic resources by providing an open critical evaluation of broadly beneficial endeavors similar to that outlined here.

General scientific implications and limitations

While publication is clearly a very useful step, this project also needs to adopt a format suitable for continuous updating, to keep up with the increasing amount of

available single cell data. These data are likely to keep growing at even faster rate, considering that comberon identification can be extended to other cell types, go beyond receptors and include more species to measure conservation.

We have only investigated the more general properties of receptor comberons in endothelial cells. For example, some receptors might belong to more than one comberon. A first step in this direction is provided by the results we present on tissue-specific receptor analysis, but there might be comberons associated with different cell states within a tissue, including disease states. More data will certainly improve the precise identification of the receptors included in the comberons, especially at the organ and tissue-specific level.

It is also important to promote scientific replication, not only internally as we have done here, but also externally. We facilitate this by sharing our software code in an organized format.

Conclusions

This paper demonstrates the existence of evolutionarily conserved combinations of receptors expressed in endothelial cells, which we call comberons. Addressing the pharmacological and socio-economic implications of this discovery would benefit from the integrated efforts of scientists from multiple disciplines.

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health (No. R01GM122085). We are very grateful for the statistical feedback received from Michael G. Walker and Karen Messer.

COMPETING INTERESTS

CP and GP own shares of Salgomed, Inc.

REFERENCES

1. Chung AS, and Ferrara N. Developmental and pathological angiogenesis. *Annual review of cell and developmental biology.* 2011;27:563-84.
2. Potente M, Gerhardt H, and Carmeliet P. Basic and therapeutic aspects of angiogenesis. *Cell.* 2011;146(6):873-87.
3. Yancopoulos GD, Davis S, Gale NW, Rudge JS, Wiegand SJ, and Holash J. Vascular-specific growth factors and blood vessel formation. *Nature.* 2000;407(6801):242-8.
4. Apte RS, Chen DS, and Ferrara N. VEGF in Signaling and Disease: Beyond Discovery and Development. *Cell.* 2019;176(6):1248-64.
5. Ferrara N, and Adamis AP. Ten years of anti-vascular endothelial growth factor therapy. *Nature reviews Drug discovery.* 2016;15(6):385-403.

6. Feala JD, Cortes J, Duxbury PM, Piermarocchi C, McCulloch AD, and Paternostro G. Systems approaches and algorithms for discovery of combinatorial therapies *WIREs Syst Biol Med.* 2010(2):181-93.
7. Kang Y, Hodges A, Ong E, Roberts W, Piermarocchi C, and Paternostro G. Identification of Drug Combinations Containing Imatinib for Treatment of BCR-ABL+ Leukemias. *PloS one.* 2014;9(7):e102221.
8. Ong E, Szedlak A, Kang Y, Smith P, Smith N, McBride M, et al. A scalable method for molecular network reconstruction identifies properties of targets and mutations in acute myeloid leukemia. *Journal of computational biology : a journal of computational molecular cell biology.* 2015;22(4):266-88.
9. Feala JD, Cortes J, Duxbury PM, McCulloch AD, Piermarocchi C, and Paternostro G. Statistical properties and robustness of biological controller-target networks. *PloS one.* 2012;7(1):e29374.
10. Jacob F, Perrin D, Sanchez C, and Monod J. The Operon: a group of genes with the expression coordinated by an operator. *Comptes rendus hebdomadaires des seances de l'Academie des sciences.* 1960;250:1727-9.
11. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. *eLife.* 2017;6.
12. HuBmap C. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature.* 2019;574(7777):187-92.
13. Nurse P. *What Is Life?: understand biology in five steps.* Scribe Publications Pty Limited; 2020.
14. Szedlak A, Smith N, Liu L, Paternostro G, and Piermarocchi C. Evolutionary and Topological Properties of Genes and Community Structures in Human Gene Regulatory Networks. *PLoS computational biology.* 2016;12(6):e1005009.
15. Karagiannis P, Takahashi K, Saito M, Yoshida Y, Okita K, Watanabe A, et al. Induced Pluripotent Stem Cells and Their Use in Human Models of Disease and Development. *Physiological reviews.* 2019;99(1):79-114.
16. Ferrara N, Gerber HP, and LeCouter J. The biology of VEGF and its receptors. *Nature medicine.* 2003;9(6):669-76.
17. Ferrara N, and Kerbel RS. Angiogenesis as a therapeutic target. *Nature.* 2005;438(7070):967-74.
18. Franzén O, Gan L-M, and Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database.* 2019;2019.
19. Franzén O, and Björkegren JLM. alona: a web server for single-cell RNA-seq analysis. *Bioinformatics.* 2020;36(12):3910-2.
20. Domanskyi S, Szedlak A, Hawkins NT, Wang J, Paternostro G, and Piermarocchi C. Polled Digital Cell Sorter (p-DCS): Automatic identification of hematological cell types from single cell RNA-sequencing clusters. *BMC bioinformatics.* 2019;20(1):369.
21. Domanskyi S, Hakansson A, Bertus T, Paternostro G, and Piermarocchi C. Digital Cell Sorter (DCS): a cell type identification, anomaly detection, and Hopfield landscapes toolkit for single-cell transcriptomics. *PeerJ.* 2021:9:e10670.
22. Voigt AP, Mulfaul K, Mullin NK, Flamme-Wiese MJ, Giacalone JC, Stone EM, et al. Single-cell transcriptomics of the human retinal pigment epithelium and choroid in health and macular degeneration. *Proc Natl Acad Sci U S A.* 2019;116(48):24100-7.

23. Ward Jr JH. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association.* 1963;58(301):236-44.
24. Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P, et al. Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. *Cell systems.* 2018;6(4):484-95 e5.
25. Kong DH, Kim MR, Jang JH, Na HJ, and Lee S. A Review of Anti-Angiogenic Targets for Monoclonal Antibody Cancer Therapy. *International journal of molecular sciences.* 2017;18(8).
26. Wooldridge JM. Quasi-likelihood methods for count data. *Handbook of applied econometrics.* 1999;2:352-406.
27. Silva JS, and Tenreyro S. The log of gravity. *The Review of Economics and statistics.* 2006;88(4):641-58.
28. Efron B, and Hastie T. *Computer Age Statistical Inference.* Cambridge University Press; 2016.
29. Soltis PS, and Soltis DE. Applying the bootstrap in phylogeny reconstruction. *Statistical Science.* 2003:256-67.
30. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *evolution.* 1985;39(4):783-91.
31. Van Noorden R, Maher B, and Nuzzo R. The top 100 papers. *Nature News.* 2014;514(7524):550.
32. Good P. *Permutation tests: a practical guide to resampling methods for testing hypotheses.* Springer Science & Business Media; 2013.
33. Manly BF. *Randomization, bootstrap and Monte Carlo methods in biology.* CRC press; 2006.
34. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun.* 2019;10(1):1523.
35. Domanskyi S, Piermarocchi C, and Mias GI. PyIOmica: longitudinal omics analysis and trend identification. *Bioinformatics.* 2019;36(7):2306-7.
36. Komili S, and Silver PA. Coupling and coordination in gene expression processes: a systems biology view. *Nature reviews Genetics.* 2008;9(1):38-48.
37. Latchman D. *Gene Control, Second Edition.* CRC Press LLC; 2015.
38. Wagner A. *Robustness and evolvability in living systems.* Princeton: Princeton University Presss; 2005.
39. Nowak MA, Boerlijst MC, Cooke J, and Smith JM. Evolution of genetic redundancy. *Nature.* 1997;388(6638):167-71.
40. Jacob F. Evolution and tinkering. *Science.* 1977;196(4295):1161-6.
41. Ohno S. *Evolution by gene duplication.* Springer Science & Business Media; 2013.
42. Dawkins R. *Climbing mount improbable.* WW Norton & Company; 1997.
43. Sakurai A, Gavard J, Annas-Linhares Y, Basile JR, Amornphimoltham P, Palmby TR, et al. Semaphorin 3E initiates antiangiogenic signaling through plexin D1 by regulating Arf6 and R-Ras. *Molecular and cellular biology.* 2010;30(12):3086-98.
44. Yang WJ, Hu J, Uemura A, Tetzlaff F, Augustin HG, and Fischer A. Semaphorin-3C signals through Neuropilin-1 and PlexinD1 receptors to inhibit pathological angiogenesis. *EMBO Mol Med.* 2015;7(10):1267-84.

45. Sturtzel C, Testori J, Schweighofer B, Bilban M, and Hofer E. The transcription factor MEF2C negatively controls angiogenic sprouting of endothelial cells depending on oxygen. *PloS one.* 2014;9(7):e101521.
46. Almen MS, Nordstrom KJ, Fredriksson R, and Schioth HB. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC biology.* 2009;7:50.
47. Antibody S. Antibody therapeutics approved or in regulatory review in the EU or US. https://www.antibodysociety.org/resources/approved-antibodies/.
48. Strohl WR. Current progress in innovative engineered antibodies. *Protein & cell.* 2018;9(1):86-120.
49. Singh S. *The code book.* New York: Doubleday; 1999.
50. Gomes T, and Teichmann SA. An antiviral response beyond immune cells. *Nature.* 2020;583(7815):206-7.
51. Krausgruber T, Fortelny N, Fife-Gernedl V, Senekowitsch M, Schuster LC, Lercher A, et al. Structural cells are key regulators of organ-specific immune responses. *Nature.* 2020;583(7815):296-302.
52. Morens DM, and Fauci AS. Emerging Pandemic Diseases: How We Got to COVID-19. *Cell.* 2020;182(5):1077-92.
53. Ackermann M, Verleden SE, Kuehnel M, Haverich A, Welte T, Laenger F, et al. Pulmonary Vascular Endothelialitis, Thrombosis, and Angiogenesis in Covid-19. *The New England journal of medicine.* 2020;383(2):120-8.
54. Blanco-Melo D, Nilsson-Payant BE, Liu WC, Uhl S, Hoagland D, Moller R, et al. Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell.* 2020;181(5):1036-45 e9.
55. Group RC, Horby P, Lim WS, Emberson JR, Mafham M, Bell JL, et al. Dexamethasone in Hospitalized Patients with Covid-19 - Preliminary Report. *The New England journal of medicine.* 2020.
56. Scannell JW, Blanckley A, Boldon H, and Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature reviews Drug discovery.* 2012;11(3):191-200.
57. National Academies of Sciences E, and Medicine. *Open Science by Design: Realizing a Vision for 21st Century Research.* Washington (DC): National Academies Press (US); 2018.
58. Palca J. The genome project: life after Watson. *Science.* 1992;256(5059):956-8.
59. Begley CG, and Ellis LM. Drug development: Raise standards for preclinical cancer research. *Nature.* 2012;483(7391):531-3.
60. Prinz F, Schlange T, and Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery.* 2011;10(9):712.
61. Williams HL. Intellectual property rights and innovation: Evidence from the human genome. *Journal of Political Economy.* 2013;121(1):1-27.
62. Murray F, Aghion P, Dewatripont M, Kolev J, and Stern S. Of mice and academics: Examining the effect of openness on innovation. *American Economic Journal: Economic Policy.* 2016;8(1):212-52.
63. Azoulay P, Graff Zivin JS, and Manso G. Incentives and creativity: evidence from the academic life sciences. *The RAND Journal of Economics.* 2011;42(3):527-54.
64. Pledge G. The Giving Pledge. https://givingpledge.org/.

65. Institute for Policy Studies. Gilded Giving 2020. https://inequality.org/great-divide/gilded-giving-2020/.
66. O'Clery C. *The Billionaire Who Wasn't: How Chuck Feeney Secretly Made and Gave Away a Fortune.* PublicAffairs; 2013.
67. Cho A. With to-do list checked off, U.S. physicists ask, 'What's next?'. *Science.* 2020;370(6513):156-7.
68. Maxmen A. Untapped potential: More US labs could be providing tests for coronavirus. https://www.nature.com/articles/d41586-020-01154-6.